Bayesian approach to outlier detection in multivariate normal samples and linear models

By

Alexandre Varbanov,
School of Statistics, University of Minnesota

Technical Report No. 614
June 26, 1996

Abstract

Chaloner and Brant (1988) propose a Bayesian method for identifying outliers in univariate linear models. This paper presents an approach generalizing their idea to multivariate normal samples and multivariate linear models. The posterior distribution of the squared norm of the realized errors is used for outlier identification. Bayes factors are used for examining whether or not an observation is an outlier.

Some key words: Bayes factors; Multivariate linear models; Outlier detection; Posterior distribution.

1 Introduction

The detection of outliers is important not only for univariate data but also for multivariate samples. Outliers are observations that do not follow the pattern of the majority of the data and show extremeness relative to some basic model. Barnett and Lewis (1994, Ch.7) point out that the idea of extremeness arises from some form of "ordering" the data. While it is straightforward to define extremeness in a univariate sample, it is not so trivial for multivariate data. Barnett and Lewis (p. 270) argue that one needs to "adopt an appropriate sub-ordering principle as a basis for expressing extremeness of observations". Barnett (1976) considers the role of sub-ordering principles in multivariate problems and classifies them in four types: reduced, marginal, partial and conditional. The reduced sub-ordering is almost the only principle that has been used for multivariate outlier detection. This principle is applied by ordering a sample of p-dimensional observations $Y_1, \ldots, Y_n$ in terms of the values of a univariate statistic R(Y). Then an observation Y_i could be suspected as an outlier, if

$$R(Y_i) = \max\{R(Y_j),\ j = 1, ..., n\}.$$

Furthermore, it could be declared as an outlier, if $R(Y_i)$ is unreasonably far from the center of its distribution under the basic model of consideration.

In the literature for frequentist methods for multivariate outlier detection most commonly used statistics take the form:

$$R(Y_j, \eta, V) = (Y_j - \eta)^T V^{-1} (Y_j - \eta). \tag{1}$$

Here, η represents the location of the sample or the popoulation and V is a measure of the sample or population variation. If the mean μ and the covariance Σ of the underlying distribution are known, then (1) becomes:

$$R(Y_j, \mu, \Sigma) = (Y_j - \mu)^T \Sigma^{-1} (Y_j - \mu). \tag{2}$$

Barnett and Lewis (1994, p.272) show that under normality, (2) "has a substantial practical appeal in terms of probability density ellipsoids and ... also has much broader statistical support".

Usually μ and Σ are unknown and they are substituted in (2) with their sample estimates - the sample mean and the sample covariance matrix. These estimates could be affected by outliers and some authors suggest using robust estimates of μ and Σ (Campbell, 1980, Rousseeuw and Van Zomeren, 1990).

The only Bayesian approach for outlier detection in multivariate samples is due to Guttman (1973). The approach assumes that the underlying distribution is $N(\mu, \Sigma)$ and there is one observation which comes from $N(\mu + \alpha, \Sigma)$. Guttman suggests using the posterior distribution

of α to detect outliers. A weight, c_j , is attached to the j-th observation, where the c_j's are inversely proportional to the determinant of the sample covariance matrix of all observations but Y_j , raised to the power $\frac{1}{2}n - 1$. According to Guttman (p. 736), "if there is an outlier , the corresponding c_j will be large, and examination of the weights will be very revealing".

Other contributions to detection of multivariate outliers are included in Gnanadesikan and Kettenring (1972), Hawkins (1980), Rousseeuw and Leroy (1987).

A Bayesian approach to detection of multivariate outliers is proposed in this paper. It is a generalization of the approach to outlier detection in univariate linear models, suggested by Chaloner and Brant (1988). They use the random errors of the model to define outliers and use the posterior distribution of the errors to detect outliers. In the Bayesian approach presented here the posterior distribution of the quantity $R(Y_j, \mu, \Sigma)$, defined in (2), is used for outlier detection. In section 2 the identification of outliers in multivariate normal samples is considered, while in section 3 the approach is applied to the multivariate linear model. Although the first case can be considered as a special case of the latter, it is presented first, because it has simpler notation and introduces the ideas.

2 Multivariate normal samples

In this section we propose a Bayesian approach to the detection of outliers in multivariate normal samples. Suppose $Y_1, \ldots, Y_n$ is a random sample of n observations from a p-dimensional normal distribution, $N(\mu, \Sigma)$. Define:

$$\begin{aligned} \epsilon_i &= \Sigma^{-\frac{1}{2}}(Y_i - \mu), \qquad\qquad i = 1, \ldots, n \\ \delta_i &= \epsilon_i^T \epsilon_i = R(Y_i, \mu, \Sigma) \end{aligned} \tag{3}$$

The ϵ_i's are independent and have known normal distribution with mean 0 and covariance matrix I (identity). We declare the i-th observation to be an outlier if the posterior probability of $\delta_i > k$ is larger than some value for an appropriate choice of k. The value of k can be chosen so that the prior probability of no outliers is large. If we choose this probability to be 0.95, then

$$0.95 = pr(\delta_i \leq k, \ for\ all\ i) = \{F_p(k)\}^n \tag{4}$$

where $F_p()$ is the distribution function of a random variable with central chi-square distribution with p degrees of freedom.

The solution of (4) for k is:

$$k = F_p^{-1}(0.95^{\frac{1}{n}}).$$

Noninformative priors for μ , Σ are used to derive the posterior distribution of the $\delta_i's$ and to compute $p_i = pr(\delta_i > k|Y)$. The p_i's can then be used to identify possible outliers. They can be ordered and the observations with the largest p_i's would be suspected as outliers. An alternative is to use Bayes factors for testing the hypotheses:

$$H_{0i} \ : \ \delta_i > k \quad (Y_i\ is\ an\ outlier)$$

$$H_{1i} \ : \ \delta_i \leq k \qquad\qquad i = 1, \ldots, n$$

The Bayes factor, B_i , for testing H_{0i} versus H_{1i} is the ratio of the posterior odds of H_{0i} to the prior odds:

$$B_i = \frac{p_i F_p(k)}{(1-p_i)\{1-F_p(k)\}}$$

Kass and Raftery (1995) suggest that values of B_i greater than 10 would suggest strong evidence for H_{0i} and values greater than 100 would show very strong evidence.

A standard choice for the priors of μ and Σ is :

$$\begin{aligned} \pi(\mu) &\propto 1 \\ \pi(\Sigma) &\propto |\Sigma|^{-\frac{p+1}{2}} \\ \pi(\mu,\Sigma) &= \pi(\mu)\pi(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}} \end{aligned}$$

(Box and Tiao, 1973 , p. 426).

Define

$$\begin{aligned} \bar{Y} &= \frac{\sum_{i=1}^{n} Y_i}{n} \\ S &= \sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})^T \end{aligned}$$

then the likelihood of the data is:

$$f(Y|\mu,\Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp[-\frac{1}{2}\{tr(\Sigma^{-1}S)+n(\bar{Y}-\mu)^T\Sigma^{-1}(\bar{Y}-\mu)\}].$$

Using the prior and the likelihood, the joint posterior distribution of (μ,Σ) is:

$$p(\mu,\Sigma|Y) \propto |\Sigma|^{-\frac{n+p+1}{2}} \exp\{-\frac{1}{2}tr(\Sigma^{-1}S)\} \exp\{-\frac{n}{2}(\mu-\bar{Y})^T\Sigma^{-1}(\mu-\bar{Y})\}$$

and hence

$$\begin{aligned} \mu|\Sigma,Y &\sim N(\bar{Y},n^{-1}\Sigma) \\ \Sigma|Y &\sim W^{-1}(S,p,n-p) \end{aligned} \tag{5}$$

where $W^{-1}(S,p,n-p)$ is the inverted Wishart distribution (Box and Tiao, 1973, p.460-464). Using (5), the conditional posterior distributions of $\epsilon_i|\Sigma,Y$ and $\delta_i|\Sigma,Y$ can be derived. The ϵ_i's are linear functions of μ , given Σ. Defining

$$\gamma_i = \Sigma^{-\frac{1}{2}}(Y_i-\bar{Y})$$

we get from (3) and (5) that

$$\epsilon_i|\Sigma, Y \sim N(\gamma_i, n^{-1}I)$$

and hence, $W_i = n\delta_i$, given Σ and Y, has a noncentral chi-square distribution with p degrees of freedom and noncentrality parameter

$$\lambda_i = n(Y_i - \bar{Y})^T \Sigma^{-1}(Y_i - \bar{Y}).$$

The above distributional results allow us to write p_i $(i = 1, \ldots, n)$ as:

$$\begin{aligned} p_i &= E_{\Sigma|Y}[pr(\delta_i > k|Y, \Sigma)] \\ &= E_{\Sigma|Y}[pr(W_i > nk|Y, \Sigma)]. \end{aligned}$$

From computational point of view it is better to write the p_i's in terms of $V = \Sigma^{-1}$:

$$p_i = E_{V|Y}[pr(W_i > nk|Y, V)]. \tag{6}$$

Here,

$$\begin{aligned} \lambda_i &= n(Y_i - \bar{Y})^T V (Y_i - \bar{Y}) \\ V|Y &\sim W(S^{-1}, p, n - p). \end{aligned}$$

Monte Carlo techniques can be used for estimating (6) for each i.

3 Multivariate linear models

In this section the approach is applied to the multivariate linear model. Let Y_i be the vector with the p response variables on the i-th sampling unit $(i = 1, \ldots, n)$. Let Y be the $n \times p$ data matrix defined by:

$$Y = \begin{pmatrix} Y_1^T \\ Y_2^T \\ \vdots \\ Y_n^T \end{pmatrix} = \left(Y^{(1)}, Y^{(2)}, \ldots, Y^{(p)}\right)$$

Here, $Y^{(j)}(j = 1, \ldots, p)$ represents n independent observations on the j-th variable. The multivariate linear model assumes that

$$Y^{(j)} = X\theta_j + \epsilon^{(j)}$$

where X is a $n \times q$ design matrix,

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix}$$

The multivariate linear model can be written also in a matrix form:

$$Y = X\Theta + E \tag{7}$$

The i-th row of the $n \times p$ matrix E contains the random errors for the Y_i :

$$E = \begin{pmatrix} \epsilon_1^T \Sigma^{\frac{1}{2}} \\ \epsilon_2^T \Sigma^{\frac{1}{2}} \\ \vdots \\ \epsilon_n^T \Sigma^{\frac{1}{2}} \end{pmatrix} = \left(\epsilon^{(1)}, \epsilon^{(2)}, \cdots, \epsilon^{(p)}\right)$$

and $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$.

It is assumed also that the ϵ_i's are independent identically distributed p-dimensional random vectors from $N(0, I)$. The model (7) can be written also as:

$$Y_i = \Theta^T X_i + \Sigma^{\frac{1}{2}} \epsilon_i \qquad\qquad i = 1, \ldots, n$$

The i-th observation Y_i is declared to be an outlier if the posterior probability of $\delta_i > k$ is larger than some value for an appropriate choice for k. The constant k could be defined in the same way as in Section 2.

Using the noninformative prior $\pi(\Theta, \Sigma) \propto |\Sigma|^{-\frac{p+1}{2}}$, the needed posterior distributions are:

$$\begin{aligned} \Theta^T|\Sigma, Y &\sim N(\hat{\Theta}^T, \Sigma \otimes (X^T X)^{-1}) \\ \Sigma|Y &\sim W^{-1}(S, p, n-q-p+1) \end{aligned}$$

where $\hat{\Theta} = (X^T X)^{-1} X^T Y$ and $S = (Y - X\hat{\Theta})^T (Y - X\hat{\Theta})$.

Let

$$\begin{aligned} \sigma_{(i)} &= X_i^T (X^T X)^{-1} X_i \\ \lambda_i &= \sigma_{(i)}^{-1} (Y_i - \hat{\Theta}^T X_i)^T \Sigma^{-1} (Y_i - \hat{\Theta}^T X_i) \end{aligned}$$

then given Σ and Y, the posterior distribution of $W_i = \frac{\delta_i}{\sigma_{(i)}}$ is noncentral chi-square with p degrees of freedom and noncentrality parameter λ_i.

So, we can write p_i $(i = 1, \ldots, n)$ as

$$p_i = E_{\Sigma|Y}\{pr(W_i > \frac{k}{\sigma_{(i)}}|Y, \Sigma)\}$$

or in terms of $V = \Sigma^{-1}$ as

$$p_i = E_{V|Y}\{pr(W_i > \frac{k}{\sigma_{(i)}}|Y, V)\}. \tag{8}$$

Here,

$$V|Y \sim W(S^{-1}, p, n-q-p+1).$$

Monte Carlo techniques can be used for estimating (8) for each i.

4 Examples

Two examples are considered in this section to support the approach for outlier detection, presented in this paper. The first example (Table 1) is the data set of Barnett and Lewis (1994, Ex.7.1, p.289).

Table 1. *Yields for plot 3 (Y_1) and plot 12 (Y_2) from 1941 to 1990*

Case	Year	Y_1	Y_2	Case	Year	Y_1	Y_2
1	1941	0.85	1.26	26	1966	1.43	2.16
2	1942	0.26	0.59	27	1967	1.31	1.48
3	1943	1.03	1.66	28	1968	1.52	1.28
4	1944	0.34	0.65	29	1969	0.72	1.87
5	1945	1.14	1.75	30	1970	1.15	1.51
6	1946	1.18	0.80	31	1971	1.50	2.94
7	1947	1.52	1.67	32	1972	1.40	1.54
8	1948	1.12	1.25	33	1973	1.24	1.27
9	1949	0.62	0.78	34	1974	1.18	1.25
10	1950	0.89	0.76	35	1975	0.91	0.55
11	1951	1.00	1.42	36	1976	1.06	1.22
12	1952	1.58	1.80	37	1977	1.20	1.21
13	1953	1.63	1.84	38	1978	1.70	1.77
14	1954	0.99	1.05	39	1979	1.26	2.27
15	1955	1.10	1.58	40	1980	0.85	1.07
16	1956	0.69	1.11	41	1981	1.47	1.95
17	1957	0.60	1.02	42	1982	2.03	1.91
18	1958	1.21	1.61	43	1983	0.99	0.84
19	1959	0.51	0.62	44	1984	1.08	1.22
20	1960	1.56	1.82	45	1985	1.58	1.65
21	1961	1.39	1.82	46	1986	0.78	1.02
22	1962	1.20	1.28	47	1987	1.39	1.71
23	1963	1.43	1.64	48	1988	1.40	1.38
24	1964	1.48	2.47	49	1989	0.60	0.75
25	1965	2.75	3.45	50	1990	0.88	0.94

For this data set:

$$n = 50, \quad p = 2, \quad k = 13.77, \quad pr(\delta_i > 13.77) = 0.001$$

$$S^{-1} = \begin{pmatrix} 0.290 & -0.171 \\ -0.171 & 0.161 \end{pmatrix}$$

The p_i's from (6) were estimated, using conditional Monte Carlo. For each i, a random matrix from the distribution $W(S^{-1}, 2, 48)$ was generated and the corresponding conditional noncentral chi-square probability was computed. This was repeated 20,000 times and the estimate for p_i was set to be equal to the average of the conditional probabilities. The only Bayes factors greater than one are: $B_{25} = 896.6$ and $B_{31} = 111.0$. The next largest Bayes factor is $B_{29} = 0.64$. So, there is strong evidence that the cases 25 and 31 are outliers. Barnett and Lewis got the same result, using a different approach for outlier detection.

The second example is the rootstock data of Rencher (1995, Table 6.2). It is a balanced data set with one factor (ROOTSTOCK) with 6 levels and 4-dimensional response (Table 2). For this data set:

$$n = 48, \quad p = 4, \quad q = 6, \quad k = 18.32,$$

$$pr(\delta_i > 18.32) = 0.0011$$

$$\sigma_{(1)} = ... = \sigma_{(48)} = 0.125$$

$$S^{-1} = \begin{pmatrix} 13.33 & -1.85 & -1.08 & 2.14 \\ -1.85 & 0.40 & -0.11 & -0.90 \\ -1.08 & -0.11 & 1.97 & -2.56 \\ 2.14 & -0.90 & -2.56 & 4.12 \end{pmatrix}$$

The p_i's were computed in a similar way as for the first example. The only Bayes factors, greater than one, are: $B_{42} = 10.48$, $B_{25} = 2.32$ and $B_{37} = 1.38$. There is strong evidence that case 42 is an outlier. Although the Bayes factors for case 25 and 37 are not very large , the two observations could be suspected as outliers too. Their posterior probability of "outlyingness" is greater than the prior probability.

Table 2. *Rootstock Data*

Case	Rootstock	Y_1	Y_2	Y_3	Y_4
1	1	1.11	2.569	3.58	0.760
2	1	1.19	2.928	3.75	0.821
3	1	1.09	2.865	3.93	0.928
4	1	1.25	3.844	3.94	1.009
5	1	1.11	3.027	3.60	0.766
6	1	1.08	2.336	3.51	0.726
7	1	1.11	3.211	3.98	1.209
8	1	1.16	3.037	3.62	0.750
9	2	1.05	2.074	4.09	1.036
10	2	1.17	2.885	4.06	1.094
11	2	1.11	3.378	4.87	1.635
12	2	1.25	3.906	4.98	1.517
13	2	1.17	2.782	4.38	1.197
14	2	1.15	3.018	4.65	1.244
15	2	1.17	3.383	4.69	1.495
16	2	1.19	3.447	4.40	1.026
17	3	1.07	2.505	3.76	0.912
18	3	0.99	2.315	4.44	1.398
19	3	1.06	2.667	4.38	1.197
20	3	1.02	2.390	4.67	1.613
21	3	1.15	3.021	4.48	1.476
22	3	1.20	3.085	4.78	1.571
23	3	1.20	3.308	4.57	1.506
24	3	1.17	3.231	4.56	1.458
25	4	1.22	2.838	3.89	0.944
26	4	1.03	2.351	4.05	1.241
27	4	1.14	3.001	4.05	1.023
28	4	1.01	2.439	3.92	1.067
29	4	0.99	2.199	3.27	0.693
30	4	1.11	3.318	3.95	1.085
31	4	1.20	3.601	4.27	1.242
32	4	1.08	3.291	3.85	1.017
33	5	0.91	1.532	4.04	1.084
34	5	1.15	2.552	4.16	1.151
35	5	1.14	3.083	4.79	1.381
36	5	1.05	2.330	4.42	1.242
37	5	0.99	2.079	3.47	0.673
38	5	1.22	3.366	4.41	1.137
39	5	1.05	2.416	4.64	1.455
40	5	1.13	3.100	4.57	1.325
41	5	1.11	2.813	3.76	0.800
42	5	0.75	0.840	3.14	0.606
43	6	1.05	2.199	3.75	0.790
44	6	1.02	2.132	3.99	0.853
45	6	1.05	1.949	3.34	0.610
46	6	1.07	2.251	3.21	0.562
47	6	1.13	3.064	3.63	0.707
48	6	1.11	2.469	3.95	0.952

5 Discussion

The use of the posterior distribution of the realized errors for residual analysis in the univariate linear models has begun in mid $70's$. It was advocated by Zellner (1975), Zellner & Moulton (1985), Chaloner & Brant (1988), Chaloner (1991) and others. The Bayesian approach, proposed in this paper, uses the posterior distribution of the squared norm of the error terms to detect outliers in the multivariate linear models. The statistic (2) seems to be without a good alternative for applying the reduced sub-ordering principle for multivariate outlier detection. By using its posterior distribution it is accounted for the uncertainty about the estimation of the unknown parameters. The approach has at least two other advantages too. It is easy for implementation and it takes only a few minutes for computing (using a PC, a FORTRAN code and the data sets from the examples).

The choice of a noninformative prior is reasonable when there is no prior information available for the parameters. Informative priors could also be used. The use of a normal prior for the mean parameters and inverse Wishart for the covariance parameters would result in similar posterior distribution for the error terms as in the paper. If other informative priors are used, then the posterior distribution of the error probably could not be available in closed form, and Markov Chain Monte Carlo techniques would be necessary for estimating the $p_i's$.

Acknowledgement

This paper was written when the author was a graduate fellow at the University of Minnesota. I am grateful to K. Chaloner for useful discussions and helping me to edit the paper. I thank B. Carlin and J. Eaton for helpful comments.

References

BARNETT, V. (1976). The ordering of multivariate data (with Discussion). *J. Roy. Statist. Soc. Ser. A*, 139, 318-54.

BARNETT, V. & LEWIS, T. (1994). *Outliers in Statistical Data.* 3rd ed. Chichester: John Wiley & Sons.

BOX, G.E.P. & TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

CAMPBELL, N. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics, 29*, 231-37.

CHALONER, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika, 78*, 637-44.

CHALONER, K. & BRANT, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika, 75*, 651-9.

GNANADESIKAN, R. & KETTENRING, J. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics, 28*, 81-124

GUTTMAN, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity - a Bayesian approach. *Technometrics, 15*, 723-38

HAWKINS, D. (1980). *Identification of Outliers.* London: Chapman and Hall.

KASS, R. & RAFTERY, A. (1995). Bayes factors. *JASA, 90*, 773-95

RENCHER, A. (1995). *Methods of Multivariate Analysis.* New York: Wiley.

ROUSSEEUW, P. & LEROY, A. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

ROUSSEEUW, P. & VAN ZOMEREN, B. (1990). Unmasking multivariate outliers and leverage points. *JASA, 85*, 633-9

ZELLNER, A. (1975). Bayesian analysis of regression error terms. *JASA, 70*, 138-44.

ZELLNER, A. & MOULTON, B. (1985). Bayesian regression diagnostics with applications to international consumption and income data. *J. Econometrics, 29*, 187-211.